HYPERPARAMETER TUNING OF TEXT-TO-SPEECH MODELS FOR EMOTIONAL SPEECH SYNTHESIS

Abhigyan Basak[1, 3], *Metta Venkata Srujan*[2, 3], *Harshal Pravin Kshirsagar*[2, 3], *M Prerana Rao*[2, 3], *Saurabh Agrawal*[4], *Gopal Kumar Agrawal*[4], *Manjula Shenoyk K*[2, 3], *Rishabh Jain*[2, 3]

[1]Department of Data Science and Computer Applications
[2]Department of Information and Communication Technology
[3]Manipal Institute of Technology,
Manipal Academy of Higher Education, Manipal, India
[4]Samsung R&D, Bangalore

ABSTRACT

In recent years, significant advancements in emotional speech synthesis have emerged, driven by deep learning-based models and vocoders. This paper conducts experiments on vital text-to-speech components, namely text-to-speech models (Tacotron 2) and vocoders (HiFi-GAN), utilizing a custom-recorded dataset featuring emotional speech samples with Indian accents across five distinct emotional states, happy, sad, neutral, angry and excited, to enhance emotional speech synthesis quality. We initially employ a Tacotron 2 model with WaveGlow vocoder and subsequently introduce the HiFiGAN vocoder. Throughout the study, we experiment with hyperparameters, activation functions and kernel sizes to optimize HiFi-GAN's performance on emotional data. This work contributes to advancing emotional speech synthesis in Indian accents and offers valuable insights into the effectiveness of different vocoder configurations.

Index Terms— Text-to-speech, emotional speech synthesis, vocoders, Tacotron 2, HiFiGAN

1. INTRODUCTION

The primary form of communication between humans is speaking. As humans, we express ourselves and our minds through speaking, sign language and gestures. It doesn't feel unnatural or artificial when there is a human touch when interacting. In today's world people on a daily basis use electronic gadgets for entertainment, work, communication etc. End-to-end text-to-speech [1, 2] systems have played a major role in minimizing the gap between the users and the electronics. It is an interactive way of communication between humans and devices and it comes with many applications as well. TTS, Text-to-speech is a technology that is used to generate audio from a text input. Emotional TTS is a branch which deals with adding emotions to the speech generated giving it a human touch in a similar fashion to how people express emotions when they communicate. The tone to TTS systems is usually constant and emotionless. The emotions can be added by changing the tempo, pitch and amplitude but the voice would sound more robotic than human. With many architectures available, this paper discusses the related experiments and results in training various state-of-the-art TTS models like Tacotron2 [3], WaveGlow [4] and HiFIGAN [5] for generating emotional text for low resource dataset. We have targeted five emotions that are Happy, Sad, Neutral, Angry and Excited. Creating speech involves not only considering linguistic and prosodic aspects but also having a deep understanding of the complex degrees of human emotions. The synthesized audio samples are available at https://abhigyanbasak248.github.io/results.github.io/icassp.html

2. RELATED WORKS

O. Kwon et. al [6] proposed an approach for emotional speech synthesis by incorporating global style tokens (GSTs) into the Tacotron2 framework. The system consists of style architecture for emotion-related style embedding and Tacotron2 for text-to-speech synthesis. From emotional speech references, the style architecture extracts style embedding vectors, while from input text sequences, the Tacotron2 encoder generates hidden linguistic feature vectors which are then combined and processed by the Tacotron2 decoder to produce mel-spectrograms and are used by the WaveNet vocoder to synthesize the emotional speech waveform. Korean speech dataset was used as corpus, encompassing happy, angry, sad, and neutral emotions. With a 95% confidence interval, the emotional evaluation results of happy (3.51 ± 0.28), sad (3.26 ± 0.17), and angry (3.58 ± 0.16) proved the relevance of the model in generating emotional speech. To make the TTS more expressive, acoustic condition modeling and sentiment analysis models based on Fastspeech2 [7] are proposed by Y. Feng et.al [8]. Two acoustic encoders are employed to ex-

tract utterance-level and phoneme-level vectors from the target speech, and a sentiment analysis model is used to derive objective sentiment features from the text, which are then expanded and fused with the acoustic model's output vector. S. K. Nithin et. al [9] evaluated the performance of two neural TTS models, Tacotron2 and Fastspeech2, in synthesizing emotional speech through Transfer Learning. Initially trained on LJSpeech dataset, both models are fine-tuned using emotional speech data from ESD. Emotion-specific models are employed to generate speech samples, achieving a classification accuracy of 79% for Fastspeech2 and 90% for Tacotron2, as assessed by the ScSer speech emotion recognition model [10].

3. DATASET

In this study, we have built a comprehensive emotional speech dataset uniquely tailored to Indian accents. This dataset encompasses a broad spectrum of human emotions: happiness, sadness, anger, neutrality, and excitement with average audio duration being 4-5 seconds. This rich dataset serves as a foundational resource for advancing the field of emotional speech synthesis and understanding the intricacies of emotional expression within the context of Indian accents, setting the stage for more nuanced and culturally relevant speech synthesis systems.

Table 1. Dataset Details

Emotion	Samples
Happy	1500
Sad	1600
Neutral	1800
Excitement	1800
Angry	500

4. EXPERIMENTS

4.1. Tacotron 2

Tacotron 2 is a fully neural TTS system that combines a sequence-to-sequence recurrent network with attention to predict mel spectrograms. It employs a two-step process: first, a text encoder converts input text into a sequence of numerical embeddings. Next, a recurrent neural network-based decoder generates a mel spectrogram, representing audio frequencies over time. This spectrogram is then converted into audible speech using a vocoder. The vocoder of choice here is WaveGlow.

We extended the capabilities of Tacotron2 by integrating several key components to enhance the quality and expressiveness of generated speech. We incorporated a dedicated prosody prediction network to capture prosodic information within the mel-spectrograms, thereby improving the intonation and rhythmic aspects of synthesized speech.

We also experimented with many hyperparameters of Tacotron2 to finetune it to our dataset using MOS Score as the metric. The experiments on the Learning Rate and Gating Threshold are given in Table 2.

Table 2. Experimentation with Learning Rate and Gating Threshold of Tacotron2

Learning Rate	Gate Threshold	MOS
$1e-4$	0.5	1.4
$1e-5$	0.9	2.2
	0.7	2.4
	0.5	2.7
	0.3	2.5
	0.1	2.1
$1e-6$	0.5	3
	0.4	2.4
	0.3	2.3

Our first experimentation revealed that Learning Rate of 1e-6 and Gating Threshold of 0.5 gave us the best results. We further experimented with Weight Decay and Dropouts, keeping the Gating Threshold as 0.5, which are listed in Table 3.

Table 3. Experimentation with Weight Decay and Dropouts of Tacotron2

Learning Rate	Weight Decay	Dropout	MOS
$1e-4$	$1e-6$	0	1.2
	$1e-6$	0.1	1.4
$1e-5$	$1e-5$	0.1	2.6
	$1e-6$	0.1	2.7
	$1e-7$	0.1	2.4
$1e-6$	$1e-5$	0	2.3
	$1e-6$	0.1	3

This experimentation further revealed that Weight Decay of 1e-6 and Dropout of 0.1 gives us the best MOS Score.

4.2. Waveglow

Waveglow is a flow based generative model used for speech synthesis. It takes in a sequence of mel-spectrogram as its input. It uses the information from these for faster audio synthesis with highly efficient results. The waveglow model uses 12 coupling layers and 12 invertible 1x1 convolutions. As Glow was successful in generating high quality realistic images while using a simple structure of invertible 1x1 convolution. Waveglow was integrated with a similar idea based of the outstanding results of Glow. The invertible 1x1 convolution allows easy computation in both forward and backward

transformations minimizing the issues while training. Waveglow after training understands the complexities and dependencies and effectively maps the mel-spectrograms and waveforms.

We employed a pretrained WaveGlow model trained on LJ Speech dataset to enhance Tacotron2 experiments. Our custom-trained Tacotron2 provided mel-spectrograms for WaveGlow inference. A denoising step post WaveGlow ensured clean audio output. Text prompts were fed into Tacotron2 for generating mel spectrograms, which were then processed by WaveGlow to produce high-quality audio waveforms.

4.3. HiFi-GAN

HiFi-GAN is a speech synthesis model that focuses on efficiently producing high-fidelity speech waveforms from mel-spectrograms. It incorporates several technical innovations to enhance sample quality and synthesis efficiency and employs a generator with a multi-receptive field fusion (MRF) module, enabling it to capture diverse patterns in audio data effectively. Moreover, the model introduces a multi-period discriminator (MPD) to specifically handle sinusoidal signals with various periods, a crucial factor in generating realistic speech audio. To enhance training stability and audio quality, HiFi-GAN also utilises a mel-spectrogram loss.

4.3.1. Kernel Experiments

We started with kernel experimentation, which aimed to optimise the performance of the resblock and upsample kernels in our deep learning model. The original resblock kernel configuration, denoted as 3,7,11, yielded successful outcomes. However, we observed that altering the order of dimensions while maintaining the same values did not produce desirable results as the model failed to generate the expected output. Similarly, other dimension permutations, such as 7,5,3, 11,7,3, 11,5,7, 5,7,11, and 5,7,3, also exhibited the same issue, indicating the sensitivity of the resblock kernel dimensions to their order.

Modifications to the upsample kernel dimensions also did not significantly affect the model's performance. The original upsample kernel configuration, 4,4,16,16, successfully generated the desired outputs. Alterations in the dimension order, such as 16,16,4,4 and 4,16,16,16, did not lead to a substantial improvement or deterioration in results. Overall, our kernel experimentation highlights the importance of selecting appropriate dimensions for resblock kernels while suggesting that upsample kernel configurations can be relatively stable in achieving the desired outcomes in our deep learning model.

4.3.2. Activation Function Experiments

In this section, we present the results of our experiments with the activation functions in the hidden and output layers. We experimented with the activation function of the output layer with Sigmoid, SiLU, Softsign and Leaky ReLU. However, we observed that they didn't perform as well as the original Tanh activation function. We have listed our experiments in Table 4 and the loss curves in Figure 1 and 2.

Table 4. Output Layer Experimentation on HiFi-GAN

Activation Function	MOS
Tanh	4.35
Sigmoid	4
SiLU	3.9
Softsign	2
Leaky ReLU	2.5

We experimented with the hidden layer activation function from the original Leaky ReLU to ELU, CELU, SELU, GELU, HardTanh and Softplus. However, we observed that none of the new activation functions provided better results than LeakyReLU. We have listed our experiments in Table 5 and the loss curves in Figure 3 and 4.

Table 5. Hidden Layer Experimentation on HiFi-GAN

Activation Function	MOS
Leaky ReLU	4.16
ELU	2.5
CELU	3.1
SELU	2.2
GELU	2.9
HandTanh	1.8
Softplus	1.2

5. CONCLUSION

In conclusion, our research in emotional speech synthesis with Indian accents has led to valuable findings. We created a dataset with five emotions and trained the Tacotron 2 model with the WaveGlow vocoder, which enhanced Indian-accented emotional speech synthesis. We then integrated the HifiGAN vocoder into our pipeline and conducted various experiments with loss functions and kernel sizes, revealing varying results. This work contributes to emotional speech synthesis for low resource dataset and underscores the importance of adapting these technologies to diverse linguistic and cultural contexts. Our dataset, methods, and experimental insights offer valuable resources for future research. We move closer to achieving emotionally expressive, culturally nuanced, and natural speech synthesis systems for Indian accents and beyond, bridging the gap between artificial and human-generated speech.

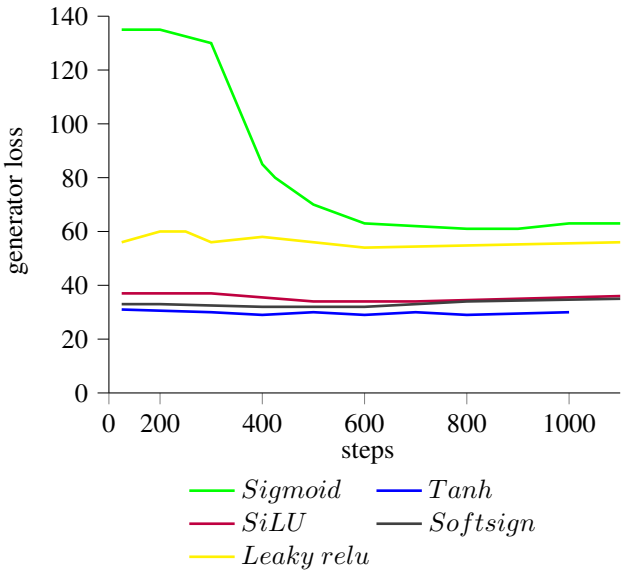

Fig. 1. Generator Loss for experiments with output layer

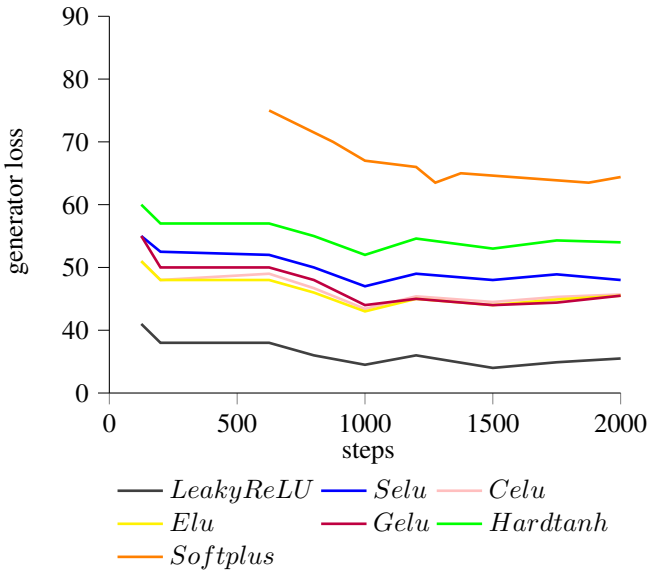

Fig. 3. Generator Loss for experiments with hidden layers

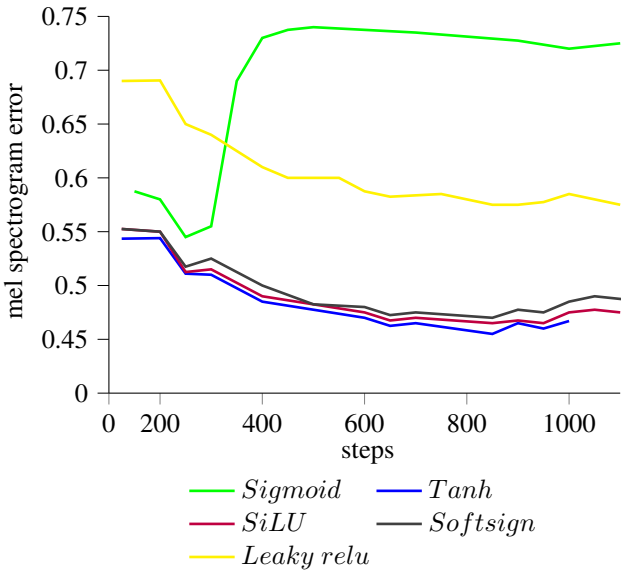

Fig. 2. Mel Spectrogram Error for experiments with output layer

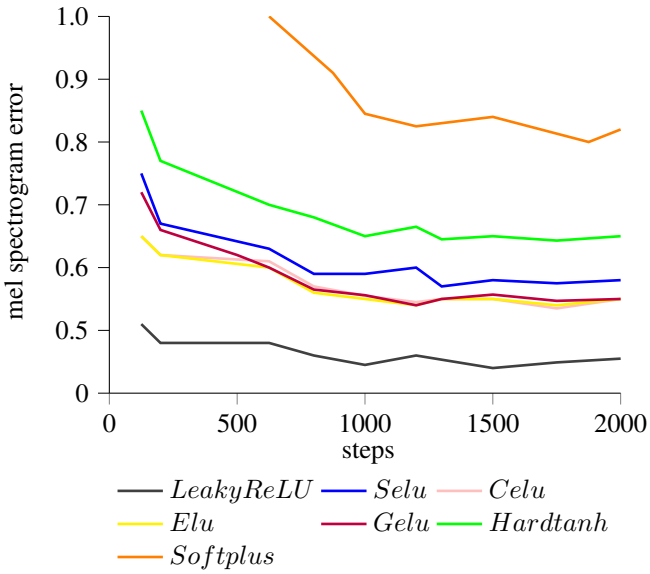

Fig. 4. Mel Spectrogram Error for experiments with hidden layers

6. REFERENCES

[1] Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, et al. Tacotron: Towards end-to-end speech synthesis. *arXiv preprint arXiv:1703.10135*, 2017.

[2] Sang-Hoon Lee, Hyun-Wook Yoon, Hyeong-Rae Noh, Ji-Hoon Kim, and Seong-Whan Lee. Multi-spectrogan: High-diversity and high-fidelity spectrogram generation with adversarial style combination for speech synthesis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13198–13206, 2021.

[3] Jonathan Shen, Ruoming Pang, Ron J Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj Skerrv-Ryan, et al. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 4779–4783. IEEE, 2018.

[4] Ryan Prenger, Rafael Valle, and Bryan Catanzaro. Waveglow: A flow-based generative network for speech synthesis. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3617–3621. IEEE, 2019.

[5] Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. *Advances in Neural Information Processing Systems*, 33:17022–17033, 2020.

[6] Ohsung Kwon, Inseon Jang, ChungHyun Ahn, and Hong-Goo Kang. Emotional speech synthesis based on style embedded tacotron2 framework. In *2019 34th International Technical Conference on Circuits/Systems, Computers and Communications (ITC-CSCC)*, pages 1–4. IEEE, 2019.

[7] Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. Fastspeech 2: Fast and high-quality end-to-end text to speech. *arXiv preprint arXiv:2006.04558*, 2020.

[8] Ying Feng, Pengfei Duan, Yunfei Zi, Yaxiong Chen, and Shengwu Xiong. Fusing acoustic and text emotional features for expressive speech synthesis. In *2022 IEEE International Conference on Multimedia and Expo (ICME)*, pages 01–06. IEEE, 2022.

[9] SK Nithin and Jay Prakash. Emotional speech synthesis using end-to-end neural tts models. In *2022 18th International Computer Engineering Conference (ICENCO)*, volume 1, pages 1–7. IEEE, 2022.

[10] Varun Sai Alaparthi, Tejeswara Reddy Pasam, Deepak Abhiram Inagandla, Jay Prakash, and Pramod Kumar Singh. Scser: Supervised contrastive learning for speech emotion recognition using transformers. In *2022 15th international conference on human system interaction (HSI)*, pages 1–7. IEEE, 2022.